DOCUMENT RESUME

ED 271 488 TM 860 386

AUTHOR Macpherson, Colin R.; Rowley, Glenn L.
TITLE An Empirical Study of the Properties of Two Estimates of Decision-Consistency Used with Two Types of Teacher-Constructed Classroom Tests.
PUB DATE Apr 86
NOTE 25p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (San Francisco, CA, April 16-20, 1986).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Classroom Research; Comparative Testing; *Criterion Referenced Tests; Cutting Scores; Decision Making; Error of Measurement; Estimation (Mathematics); Higher Education; *Mastery Tests; *Mathematical Formulas; Multiple Choice Tests; Sample Size; *Teacher Made Tests; Test Bias; Test Construction; *Test Format; *Test Reliability
IDENTIFIERS Test Specifications

ABSTRACT

Teacher-made mastery tests were administered in a classroom-sized sample to study their decision consistency. Decision-consistency of criterion-referenced tests is usually defined in terms of the proportion of examinees who are classified in the same way after two test administrations. Single-administration estimates of decision consistency were used in this study. First, it was shown that the Model 2 estimate of Marshall and Serlin's mean split-half coefficient of agreement was equivalent to the Subkoviak estimate of agreement. Ten 20-item criterion-referenced tests were constructed by 10 teachers, following strict criterion-referenced test specifications. Another 10 tests were constructed by another set of teachers, based on loosely-specified sets of objectives. These were called unit tests. It was found that the two estimation models examined produced estimates which were generally acceptable with teacher-constructed tests and classroom-sized samples, but the Model 2 estimate was marginally preferable in that it was generally less biased than Model 1 estimates. There was no systematic difference between the criterion referenced tests and the unit tests with respect to decision-consistency accuracy. (GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED271488

AN EMPIRICAL STUDY OF THE PROPERTIES OF TWO ESTIMATES OF DECISION-CONSISTENCY USED WITH TWO TYPES OF TEACHER-CONSTRUCTED CLASSROOM TESTS

Colin R. Macpherson
Hawthorn Institute of Education

and

Glenn L. Rowley
Monash University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

G. L. Rowley

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 860 386

Paper presented at the Annual Conference of the National Council for Measurement in Education, San Francisco, April, 1986.

1. Introduction

Concern about issues relating to reliability with criterion-referenced tests has, since Hambleton & Novick (1973), commonly focussed on the property of decision-consistency. This is usually defined in terms of the proportion of examinees who are classified in the same way after two administrations, either of the same test, or of parallel forms of a test. The suggestion by Hambleton and Novick (and detailed by Swaminathan, Hambleton & Algina, 1974) was that the index

$$p_o = \sum_{k=1}^{m} p_{kk}$$

would serve as a definition of CRT reliability, where m is the number of mastery levels (usually two) and p_{kk} is the proportion of examinees assigned to the kth mastery level on both tests. For the situation under consideration here, where the role of the test is the assignment of examinees into just two mastery levels ("masters" and "non-masters"), the index becomes

$$p_o = p_{11} + p_{22}$$

Thus p_o is in effect the proportion of decisions that are in agreement, and p_o is referred to as a coefficient of agreement, or of decision-consistency. Mercifully, the use of the term "reliability" to describe p_o is quite rare, nowadays.

Conceptually, the idea is attractive, but in practice the requirement of two test administrations makes such a coefficient unlikely to gain wide usage. As a result, three major thrusts have developed in estimating decision-consistency from a single test administration (Huynh, 1976; Subkoviak, 1976; Marshall, 1976). All of the methods of estimation rely on strong assumptions about the distribution of scores on an hypothetical second test - assumptions which, on the face of it, will not be met and are likely to be seriously violated by most criterion-referenced test data. In two of these methods (Subkoviak, 1976; Marshall, 1976) it is assumed that if an examinee is repeatedly tested with items from the same domain, then his or her total scores will form a binomial distribution. This derives from the assumption that an examinee has the same probability

of success on all items. With a good estimate of this probability from one test administration, the distribution of scores on an hypothetical second test administration can be estimated, and the consistency of classification decisions estimated for any chosen cut-off score. The two estimates of this probability studied in this research will be referred to as

1. Model 1, which uses the proportion correct from the items actually answered, and

2. Model 2, which uses a regression estimate based on group statistics and the linear regression estimate of true score on observed score.

Both models have been used to estimate decision consistency (although using different estimation procedures) by Subkoviak (1976) and Marshall & Serlin (1979). Model 1 and Model 2 yield different estimates of the decision-consistency one could expect had a second test been administered. The quality of these estimates has been studied in two situations:

1. with computer-simulated data, (Algina & Noe, 1978; Marshall & Serlin, 1979), and

2. with data from classically parallel norm-referenced tests, such as the SAT (e.g. Subkoviak, 1978).

Such studies, while useful as indicators, leave open the question of how well the two estimates work with real data from genuine criterion-referenced tests. And no studies could be located which looked at the adequacy of these estimates with classroom-sized samples using teacher-constructed tests. This study was designed to illuminate these questions with respect to the Model 1 and Model 2 estimates of Marshall and Serlin's (1979) coefficient. This coefficient can be shown (as will be demonstrated in Section 4 of this paper) to be numerically identical to the Subkoviak (1976) coefficient, provided that the same model is used to estimate true score.

2. Subkoviak's single-administration coefficient of decision-consistency (1976)

Where two parallel test forms are administered, Subkoviak accepted the Hambleton and Novick (1973) definition of decision-consistency as the proportion of examinees classified in the same way on the two administrations, or the probability of consistent classification averaged across examinees. To estimate this from a single test administration, Subkoviak assumed that for every examinee, the probability distribution of observed scores is binomial and therefore determined by that peron's true proportion-correct score and the number of items on the test. Given a true porportion-correct score of ζ_i for individual i, the probability of that person obtaining a score X_i greater than or equal to a given cut-off score C is given by

$$Pr\ (X_i \geq C) = \sum_{X_i = c}^{n} \binom{n}{X_i} \zeta_i^{X_i} (1 - \zeta_i)^{n-X_i}$$

where X_i is the observed score of individual i on a test of n items. From this, we see that the probability of being consistently classified on two parallel administrations is

$$p_i = \left[Pr\ (X_i \geq C)\right]^2 + \left[1 - Pr\ (X_i \geq C)\right]^2$$

This probability is averaged over examinees to yield an estimate $\hat{p}_o$ of the proportion of examinees who would be consistently classified on a parallel test.

First this requires that there be an estimate for each individual i of ζ_i, the true proportion-correct score. The two estimates used in this study, and referred to as Model 1 and Model 2 are

Model 1: $\hat{\zeta}_i = (X_i/n)$, the observed proportion-correct score for individual i, and

Model 2: $\hat{\zeta}_i = \alpha_{21} (X_i / n +) (1 - \alpha_{21}) (\bar{X} / n)$

where α_{21} is the Kuder-Richardson 21 coefficient for this set of scores, and $\bar{X}$ is the mean score of the group. This is the regression estimate of true score familiar from classical test theory.

3. The Marshall coefficient

What appears to be a quite different approach originates with the work of Marshall and Haertel (1975). Originally, they proposed to create all possible pairs of split halves from a single test of dichotomously scored items, and to compute β, the proportion of consistent decisions averaged over all possible splits and all examinees. This approach was linked to Cronbach's coefficient alpha and named coefficient beta by Marshall and Haertel. Although calculable, the coefficient described a property of a half-length test, not a full-length test, and there was no equivalent of the Spearman-Brown formula to make the appropriate adjustment.

Marshall (1976) tackled this problem as follows. On the basis of an individual's score on an n-item test, we estimate the score that person would obtain on an (hypothetical) 2n-item test if it were administered. Once this is done, the admittedly complex computational procedures of the Marshall and Haertel (1975) method can be applied. This will yield a coefficient that describes a property of a test of the appropriate length, n items.

The problem with this, however, is that even though each person's score on the n-item test is known, the distribution of scores on the non-existent 2n-item test has to be estimated. This is the same difficult faced by Subkoviak (1976) and by Huynh (1976), who had to simulate a score distribution on a hypothetical second form given the first form results, and Marshall tackled it similarly. He used the proportion-correct score X_i/n as an estimate of the true proportion-correct score ζ_i. This is Model 1, above. In later work Marshall & Serlin 1979 added the regression estimate (Model 2 above), Huynh's beta-distribution estimate, and others.

4. The equivalence of the Subkoviak and Marshall coefficients

Marshall and Serlin (1979), using simulated data, noted that, using the Model 2 estimation procedures described above, estimates of Marshall's and Subkoviak's p_0 were numerically identical. No proof was offered, but other workers in the field (Subkoviak, 1978, 1980; Wilcox, 1981) have treated the Model 1 estimate of Marshall's β as an estimate of p_0.

This apparent equivalence was verified by Macpherson (1985), who compared 200 values of the two indices using computer programs written by himself, and using data from 20 classroom tests.

Both Subkoviak and Marshall assume that the tests used to generate observed scores are simply samples of items from a large domain. In each case, the true score ζ_i of any examinee with observed score X_i is the proportion of items in the domain that the examinee would answer correctly if given the opportunity. For a single examinee, Subkoviak (1976) takes the estimate of ζ_i (whether it be Model 1 or Model 2 is immaterial here) and uses it to estimate the probability p_i of mastery on any one random sample of n items from the domain. Given this estimate, we can estimate the probability of consistent classification on any two n-item samples, given the same estimated true score. Finally, the probability of consistent classification across the entire examinee group is determined - this being the estimate of p_o.

i.e.

$$\hat{p}_o = \sum_{X=0}^{n} f_x \frac{\left[p_x^2 + (1 - p_x)^2\right]}{N}$$

where f_x = frequency of score X.

p_x = estimated true score of an examinee with observed score X, and

N = number of examinees.

In calculating the Marshall coefficient, we similarly use the observed score X on an n-item test to generate an estimate of true score ζ. Model 1 or Model 2 may be used to do this. Using this estimate, the same distributional assumptions are made to estimate the score distribution on an hypothetical 2n-item test from the same domain. Combinatoric formulae are then used to determine the probability of consistent mastery or non-mastery over every possible pair of n-item splits of the 2n-item test. But to take a 2n-item sample and to randomly split it into two n-item tests is logically equivalent to the Subkoviak procedure of taking two independent n-item samples from the domain. In computing β, Marshall uses the mean probability of consistent classification over all possible n-item splits. But on any one of those n-item splits, the probability of consistent classification is the same as if we had taken two independent n-item samples - i.e., the same as the Subkoviak probability. When we average the probabilities over all of Marshall's splits, we must get the same result as does Subkoviak.

To put it simply, Marshall's β is necessarily identical to Subkoviak's $\hat{P}_o$, provided the estimation procedures start from the same estimate of true score, or domain score. Subkoviak's procedure uses a particular set of assumptions that lead to a probability estimate based on any pair of n-item samples. Marshall's procedure uses the same set of assumptions, but determines the average probability across many pairs of n-item samples - all of these pairs giving rise to exactly the same probability of consistent classification as does the Subkoviak method.

In this study, comparisons are made between the properties of Model 1 and Model 2 estimates of decision consistency. Whether they be referred to as Marshall coefficients, or as Subkoviak coefficients is immaterial, since the two are identical. The Subkoviak procedure was used because it was computationally much faster (as would be expected given the analysis above), and the notation p_{o1} (for a Model 1 estimate of p_o) and p_{o2} (for a Model 2 estimate of p_o) will be followed.

5. Data collection

In this study, ten criterion-referenced test (CRTs), each of twenty items, were constructed by ten teachers strictly following the CRT authoring procedures laid down by Popham (1978). These tests were thought to meet the definition of a CRT about as accurately as could be reasonably expected in classroom practice. Another ten tests, based on loosely-defined sets of objectives, were constructed by another set of teachers. These are referred to as "unit tests" (UTs), and are thought to be more representative than CRTs of the type of tests commonly used in classrooms. Prior to the analysis of the data it was felt that these UTs, because they contain items based on large and vaguely defined domains, would follow the binomial assumptions of item difficulty far less accurately than would the CRTs.

Each test had a multiple-choice format, with each item having one correct answer and three distractors. Also, the scores arising from any test were used to divide students into two groups (e.g., "masters" and "non-masters") depending on how each student's score compared to a cut-off score. Both sets of tests covered a wide variety of subject-matter and a substantial range of grade levels. Such information about each of the tests is shown in Table 1.

Insert Table 1 about here

6. Method of analysis

The twenty-item tests[1] were regarded as "double-length" tests, and were used to calculate the exact value of Marshall's split-half decision-consistency coefficient (Marshall & Haertel, 1975). With the aid of a computer program, estimates of this quantity were then calculated using Model 1 and Model 2 with thirty repeated samples of ten items, treated as ten-item tests. Because the "true" value was already known, the characteristics of the two estimates were able to be assessed, with particular reference to:

(i) the amount of scatter, or random error e, about their mean values ($\overline{p_{o1}}$ or $\overline{p_{o2}}$) at each possible cut-off score, over the thirty test selections

(ii) the difference between the mean value of each estimate and the true value p_o, this being a measure of bias b, and

(iii) the "total error" E_T, which includes both of the above

For any single estimate, the relationship between these three quantities, is given by

$$E_T = b + e$$

When samples of half-tests are drawn from a population of half-tests, as happened in this study, then the statistic that generalizes the effects of these separate total errors is the "standard total error" S_{ET}. It can be shown that

$$S_{ET} = \sqrt{b^2 + S_E^2}$$

where S_E = Standard random error (i.e., the standard deviation associated with the mean estimate).

1. Except for CRT.3, which had only 18 items.

Bias alone is sometimes used as a measure of the accuracy of particular estimating techniques. However, such an approach does not take into account the random fluctuations of single estimates about their mean value. When comparing the accuracy of different estimating techniques (as is done in this study) it is important to consider both types of error; S_{ET} effectively describes the overall situation.

If S_{ET} for a particular estimator were calculated at every cut-off score then a "band" of standard total errors around the mean estimated values could be imagined. Unlike the case where standard random error (S_E) might be used, this band will always include the true coefficient value (p_o) because S_{ET} includes the effect of bias. Figure 1 provides a graphical representation of this situation for the estimator p_{o1} when used with the data from one of the unit tests.

Insert Figure 1 about here

As well as examining the errors associated with the two estimators, the extent of the errors in relation to the following properties were also investigated:

(i) the characteristics of the score distributions,

(ii) the position of the cut-off score in relation to the score distribution,

(iii) the type of test (CRT or UT),

(iv) the variance of item difficulties (which reflects the extent of departure from the binomial assumption)

The test data were also examined for any relationship between the true and estimated values of decision consistency and traditional reliability as measured by KR20.

7. Results

7.1 The behaviour of the estimates.

The values of both the "true" coefficient (p_o), and the two estimators (p_{o1} and p_{o2}), were generally lower when the cut-off score was in areas of high score density than when it was in areas of low score density. This was fully expected since it is obviously harder to consistently classify examinees when most of them have scores near the cut-off score. Such behaviour was fully demonstrated by Marshall & Haertel (1975) and Marshall & Serlin (1979). Figure 2 illustrates this general behaviour by using the same set of axes to graph the score distribution, and the coefficient values. Separate graphs of this kind produced for each of the tests used in the study showed similar characteristics.

Insert Figure 2 about here

7.2 Bias in the estimators.

The behaviour of the estimators in respect of bias can be summarized as follows:

If the distribution is unimodal with no large density at the maximum possible score, then both p_{o1} and p_{o2} show positive bias (tend to overestimate p_o) in the high score density region - p_{o2} generally showing less bias than p_{o1}. In low density areas both p_{o1} and p_{o2} show small negative bias. Any departures from this distribution type (e.g., multimodal or a significant maximum score frequency) causes p_{o2} (but not p_{o1}) to depart from this general pattern of behaviour. Even in these cases however, p_{o2} tends to give a less biased estimate.

The "errant" behaviour of p_{o2} alluded to above for some distribution types can be explained in terms of two circumstances:

(a) If the maximum possible score is gained by some examinees then, because of the nature of the binomial probability formula and the true score estimates used, p_{o2} would be expected to underestimate p_o at the maximum possible score (see Macpherson, 1985). An example of this type of behaviour of p_{o2} is shown in Figure 3.

Insert Figure 3 about here

(b) The results of this study show that p_{o2} does not reflect the characteristics of multimodal distributions quite so well as p_{o1}, and nowhere near as well as the "true" p_o. Even in these cases however, the estimated value has only a moderate degree of bias. This observation tends to reconcile the seemingly contradictory views of Subkoviak (1976), who gave theoretical reasons why the model 2 estimates should not be used with multimodal distributions, and Marshall and Serlin (1979), who found that p_{o2} did reflect the modes in such distributions except when the modes were close together.

When the realistic cut-off score ratios of 0.6, 0.7, 0.8 and 0.9 were used, (as was done for all the accuracy analyses) it was found that p_{o2} was, generally, slightly less biased than p_{o1}. This can be seen from the figures in Table 2.

Insert Table 2 about here

7.3 Accuracy in terms of total error.

When the accuracy of the estimators was considered in terms of total error (i.e., both bias and random error taken into account when single estimates are made), it was found that p_{o1} and p_{o2} were not systematically different. This can be seen from the first four columns of figures in Table 3. The figures in the last two columns however, show that p_{o2} has generally less varied total errors.

Insert Table 3 about here

7.4 Test type and estimator accuracy.

The entries in Table 3 also demonstrate that, in terms of the accuracy of decision-consistency estimates, there was no systematic difference between CRTs and UTs. Furthermore, no systematic differences could be found between the two different types of test when various statistical properties of the two groups of data were compared. In other words, after examining the general shape, skewness, modality, and maximum possible score frequency of the score distribution graphs, together with the characteristics of mean score, standard deviation, KR20 values, and departure from the binomial assumption, there was no way of distinguishing data arising from the CRTs from those arising from UTs.

7.5 Departure from the binomial assumption and accuracy of estimates.

Even though the degree of departure from the binomial assumption did not appear to be related to test type, it was felt that such a measure might be useful in predicting the accuracy of one or both estimators (since both are based on the binomial assumption). Since the fundamental binomial assumption is that all the items on a particular test have the same difficulty, it seemed reasonable to use σ_{Diff}, the standard deviation of item difficulties on a test, as the measure of departure from this assumption. However, as shown by the figures in Table 4, increasing σ_{Diff} does not appear to be systematically related to the accuracy of the estimates (as measured by the maximum standard total error over the range of cut-off score ratios 0.6, 0.7, 0.8 and 0.9). This result was surprising but it does demonstrate the robustness of the binomial assumption in this situation.

Insert Table 4 about here

7.6 Traditional reliability and the decision-consistency coefficients.

In their 1975 paper Marshall & Haertel state:

> "There seems to be a fairly high correlation across test types between α (i.e., KR20) and $\overline{\beta}$, the mean value of β over criterion levels" (p. 9)

the β referred to being the same as the true p_o used in this study. Such a relationship could not be found from the data used in this study. The product-moment correlation coefficient between mean p_o and KR20 (adjusted for half-length test) was not significantly different from zero.

Algina & Noe (1978), when investigating Subkoviak's p_o, noted that KR20 "may be a poor guide to the choice of estimator" (p. 109). Table 5 reinforces this point by listing the increasing KR20 values (adjusted by the Spearman-Brown formula for tests half as long) for the CRTs used in this investigation, against the largest standard total error (S_{ET}) in p_{o1} and p_{o2} across the same cut-off score ratios referred to above. (This table was also used in a more extensive study of whether KR20 and the proximity of the cut-off score to the mean score were related to the accuracy of p_{o1} and p_{o2}). As can be seen from this table, there appears to be no relationship between KR20 and the accuracy of either estimator.

Insert Table 5 about here

---

8. Conclusions

The results that have been presented have demonstrated the following.

1. Both estimators gave generally satisfactory results that were related as expected with the properties of the score distributions and the locations of the cut-off scores.
2. Generally, the Model 2 estimates were slightly less biased than Model 1 estimates: this remained so even in the otherwise anomalous situations where the score distributions
   (a) were multimodal, or
   (b) contained large numbers of maximum scores.

3. In terms of total error, Models 1 and 2 were not systematically different, but a point in favour of Model 2 was that it produced generally less varied total errors.

4. There was no systematic difference between CRTs and UTs with respect to the accuracy of the decision-consistency estimates.

5. For both models, the size of the standard total error was unrelated to the degree of departure of the data from the binomial assumption. The procedures seem to be robust in the face of violations of this assumption.

6. Neither the size nor the accuracy of the decision-consistency estimates obtained were systematically related to traditional reliability.

It was concluded that both models produce estimates which are generally acceptable with teacher-constructed tests and classroom-sized samples. The Model 2 estimate is preferable in that it was generally less biased than Mode 1 estimates: it had an average standard total error of 0.07 over a range of typical cut-off scores; and had a standard total error less than 0.10 in 90 percent of cases studied. The suitability of the two indices was largely independent of the statistical properties of the tests and of their mode of construction.

Finally, it is noted that neither the concept of decision consistency nor the procedures used to estimate it require that the test be truly criterion-referenced, in the sense in which the term is usually defined. The concept of decision-consistency is equally applicable and just as useful for a norm-referenced test used for selection purposes, or for a criterion-referenced test used for placement. In both cases, a decision is made based on a cutting score: the consistency with which this decision can be made is a matter for concern, regardless of whether the decision was competitive (norm-referenced), or based on a preset standard (criterion-referenced). With hindsight, this accounts for the similar characteristics of the decision-consistency estimates for the two types of test used in this study.

REFERENCES

Algina, J., & Noe, M.J. (1978). A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 15, 101-110.

Hambleton, R.K., & Novick, M.R. (1973). Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 10, 159-170.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 13, 253-264.

Macpherson, C.R. (1985). An investigation of two decision-consistency indices when used with teacher-constructed mastery tests. Unpublished Ph.D. thesis, Monash University, (Melbourne, Australia).

Marshall, J.L. (1976). The mean split-half coefficient of agreement and its relation to other test indices: A study based on simulated data. (Technical Report No. 350). Madison, WI: Wisconsin Research and Development Center for Cognitive Learning.

Marshall, J.L., & Haertel, E.H. (1975). A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement. Paper presented at the Annual Conference of the American Educational Research Association, Washington.

Marshall, J.L., & Serlin, R.C. (1979). Characteristics of four mastery test reliability indices: Influence of distribution shape and cutting score. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco.

Popham, W.J. (1978). Criterion-referenced Measurement. Englewood Cliffs, NJ: Prentice-Hall.

Subkoviak, M.J. (1976). Estimating reliability for a single admidnistration of a criterion-referenced test. Journal of Educational Measurement, 13, 265-275.

Subkoviak, M.J. (1978). Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 15, 111-116.

Subkoviak, M.J. (1980). Decision-consistency approaches. In R.A. Berk (Ed.), Criterion-referenced Measurement: The State of the Art. Baltimore: Johns Hopkins Press.

Swaminathan, H., Hambleton, R.K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 11, 263-267.

Wilcox, R.R. (1981). The single-administration estimate of the proportion of agreement of a proficiency test scored with a latent-trait structure model. Educational and Psychological Measurement, 41, 389-400.

TABLE 1

TEST CODES AND SUBJECT AREAS FOR THE CRTs AND UTs

Test Code	No. of Students Tested	Subject Area
CRT.1	29	Senior Mathematics - linear programming
CRT.2	21	Plumbing apprenticeship - air flow in drainage systems
CRT.3	31	Applied Science Nursing - electro-cardiogram interpretations
CRT.4	28	Primary - time telling with traditional clock faces
CRT.5	14	Senior Secretarial Course - wordprocessing
CRT.6	25	Senior Mathematics - absolute functions
CRT.7	30	Junior Mathematics - place value in numerals
CRT.8	26	Junior Health Education - teeth care
CRT.9	11	Senior Accounting - company accounting
CRT.10	20	Junior Mathematics - recognizing geometric figures
UT.1	26	Senior Mathematics - matrices
UT.2	31	Junior French - the use of particular types of verbs
UT.3	29	Religious Education - biblical quotes and interpretations
UT.4	27	Senior Mathematics - introductory probability
UT.5	38	Geography - Sun and rain
UT.6	27	Mathematics - permutations and combinations
UT.7	29	Junior Geography - reading and interpreting weather maps
UT.8	26	Biology - plant and animal reproduction
UT.9	22	Biology - nutrition and growth
UT.10	22	Pure Mathematics - transformations

TABLE 2

BIAS IN THE ESTIMATES OF p_{o1} AND p_{o2}

Test type	Mean absolute bias		Mean absolute percent bias		No. of times closer to Po*	
	p_{o1}	p_{o2}	p_{o1}	p_{o2}	p_{o1}	p_{o2}
CRT	0.057	0.042	8.7	6.3	13	24
UT	0.057	0.040	8.7	5.5	14	24

* Total of 40 cases for CRTs and 40 cases for UTs; equally close in 5 cases.

TABLE 3

STANDARD TOTAL ERROR IN THE ESTIMATES OF p_{o1} AND p_{o2}

Test type	No. of times S_{ET} is less		Mean of S_{ET} values		S.D. of S_{ET} values	
	p_{o1}	p_{o2}	p_{o1}	p_{o2}	p_{o1}	p_{o2}
CRT	21	19	0.070	0.068	0.055	0.040
UT	21	19	0.072	0.066	0.053	0.031

TABLE 4

THE EFFECT ON THE STANDARD TOTAL ERROR IN p_{o1} AND p_{o2} OF VIOLATIONS OF BINOMIAL ASSUMPTION

σ_{Diff}	Test	Maximum S_{ET}	
		p_{o1}	p_{o2}
0.02	CRT.3	0.067	0.037
0.03	CRT.7	0.039	0.037
0.07	UT.1	0.043	0.083
0.08	CRT.6	0.060	0.092
0.10	UT.7	0.153	0.069
0.12	CRT.10	0.060	0.077
0.12	UT.2	0.065	0.068
0.15	CRT.1	0.107	0.087
0.15	CRT.8	0.227	0.116
0.16	CRT.9	0.238	0.077
0.17	UT.3	0.119	0.099
0.17	UT.9	0.161	0.102
0.18	UT.10	0.131	0.141
0.19	CRT.4	0.105	0.090
0.21	UT.4	0.253	0.124
0.21	UT.5	0.086	0.111
0.24	UT.6	0.149	0.086
0.26	CRT.5	0.221	0.107
0.26	UT.8	0.170	0.084
0.31	CRT.2	0.159	0.105

TABLE 5

KR20 AND THE ACCURACY OF p_{o1} AND p_{o2}

Test (CRT)	Mean SCORE[1]	KR20[2]	p_{o1} MAX. S_{ET}	p_{o1} CUT-OFF	p_{o2} MAX. S_{ET}	p_{o2} CUT-OFF
3	0.96	-0.18	0.067	0.9	0.037	0.9
8	0.86	0.24	0.227	0.9	0.116	0.8
2	0.55	0.28	0.159	0.6	0.105	0.8
5	0.74	0.38	0.221	0.8	0.107	0.7
7	0.96	0.42	0.039	0.9	0.037	0.8
1	0.82	0.43	0.107	0.9	0.087	0.8
9	0.66	0.47	0.238	0.8	0.159	0.7
10	0.92	0.64	0.060	0.8	0.077	0.8
6	0.84	0.66	0.060	0.9	0.092	0.8
4	0.61	0.72	0.105	0.6	0.090	0.8

1. Mean score ratio for the 20-item test, (18 items in the case of CRT.3).

2. Adjusted for half length tests.

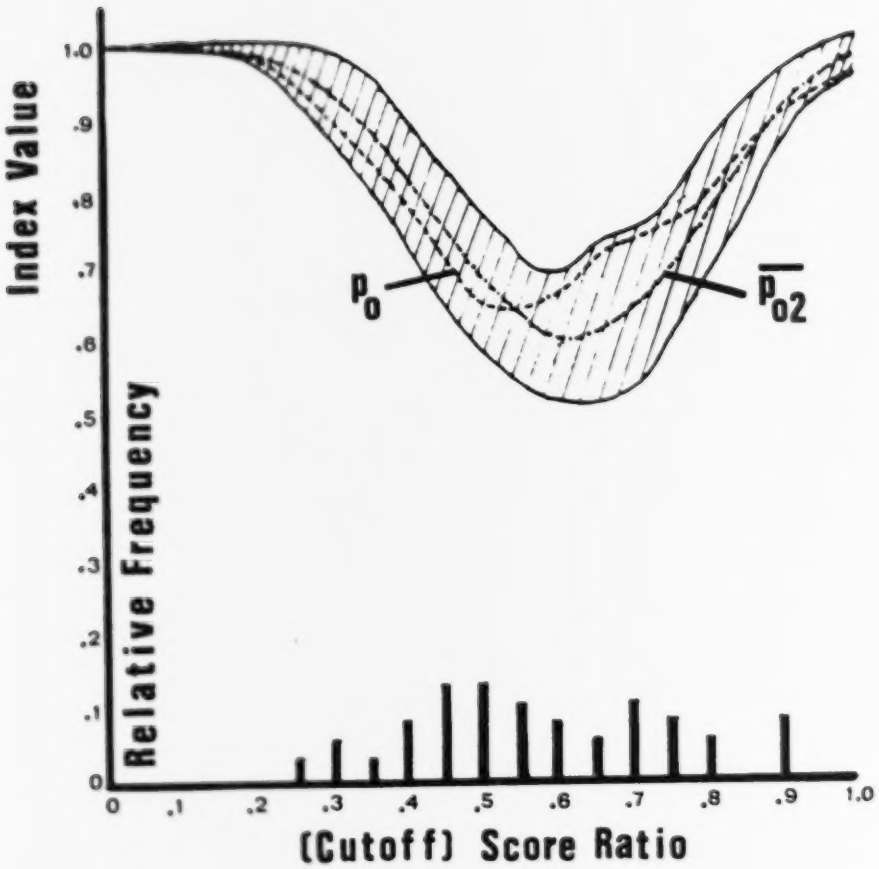
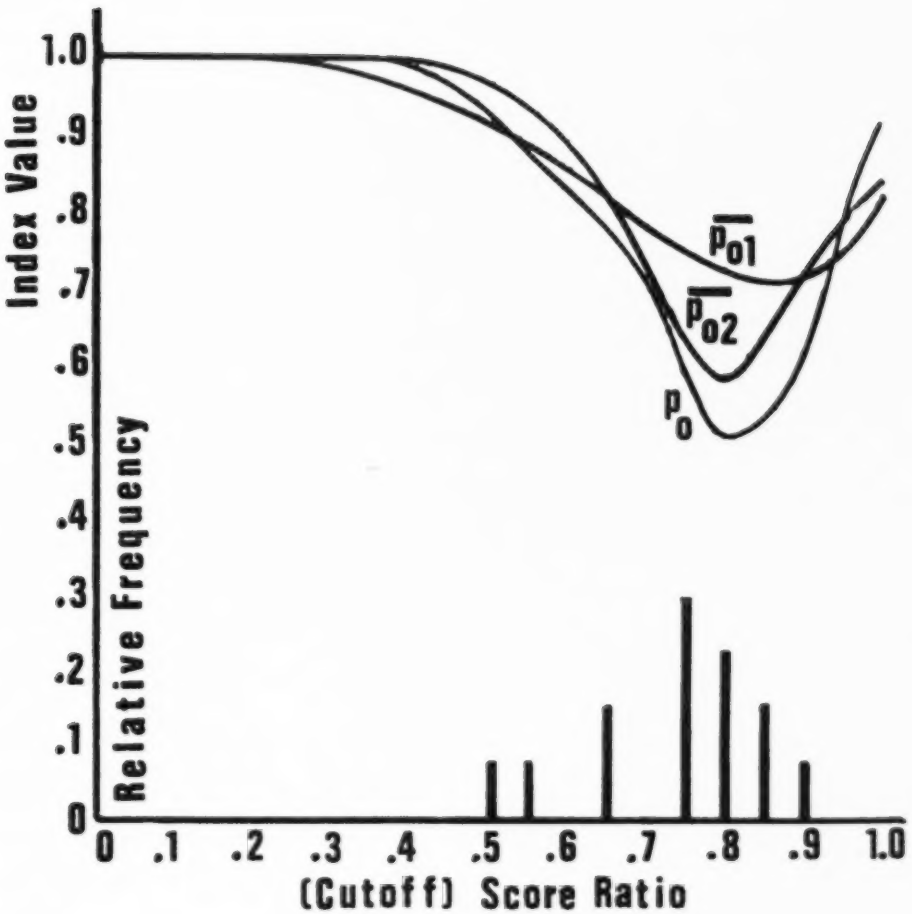


Figure 1

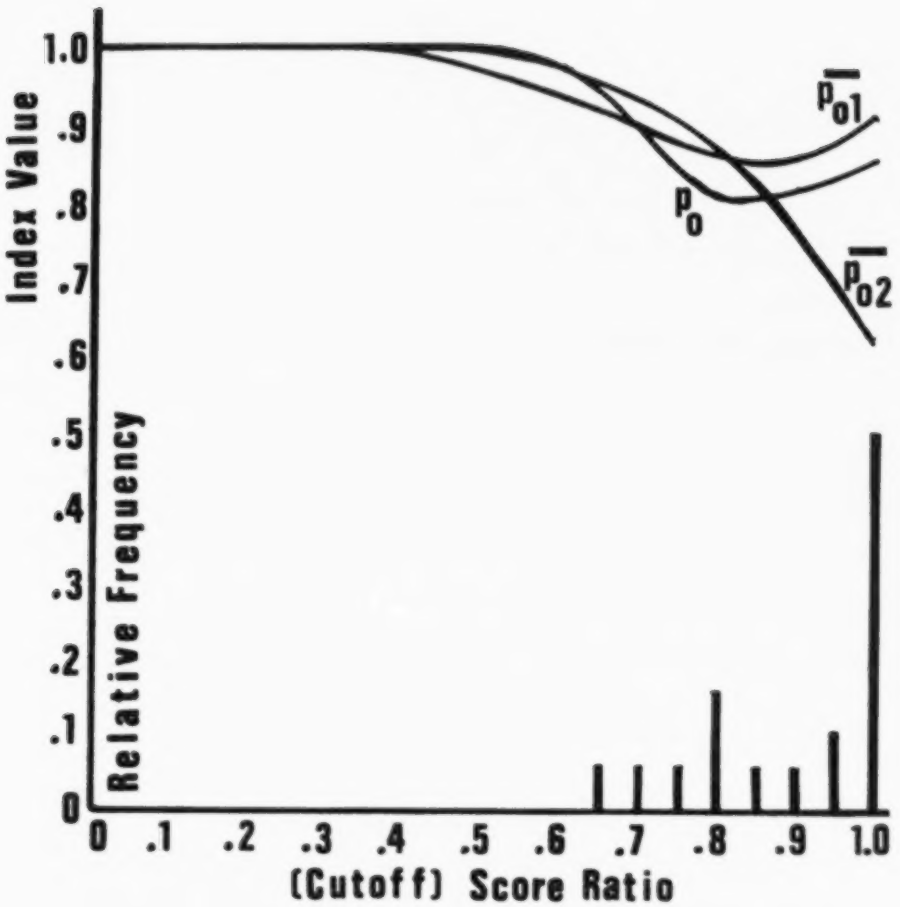
The "band" of Standard Total Error around the mean p_{o2} values for UT.5



n = 14
Mean = 0.74
S.D. = 0.11
KR20 = 0.55
KR21 = 0.28

Figure 2

The behaviour of p_o and its estimates for the test CRT.5



n = 20
Mean = 0.91
S.D. = 0.12
KR20 = 0.78
KR21 = 0.73

Figure 3

The behaviour of p_o and its estimates for the test CRT.10